Breakout Session 6:

Migration to Cloud of the Oncogenomics Next Generation Sequencing Pipelines & Databases for CCDI and Other Pediatric Cancers

Dr. Javed Khan

Senior Investigator, CCR, NCI

Migration to Cloud of the Oncogenomics Next Generation Sequencing Pipelines & Databases for CCDI and Other Pediatric Cancers

PI: Dr. Javed Khan

Current Project Personnel: Hsien-Chao Chou, Vineela Gangalapudi, Vishal Koparde, Jun Wei and Patrick Zhao

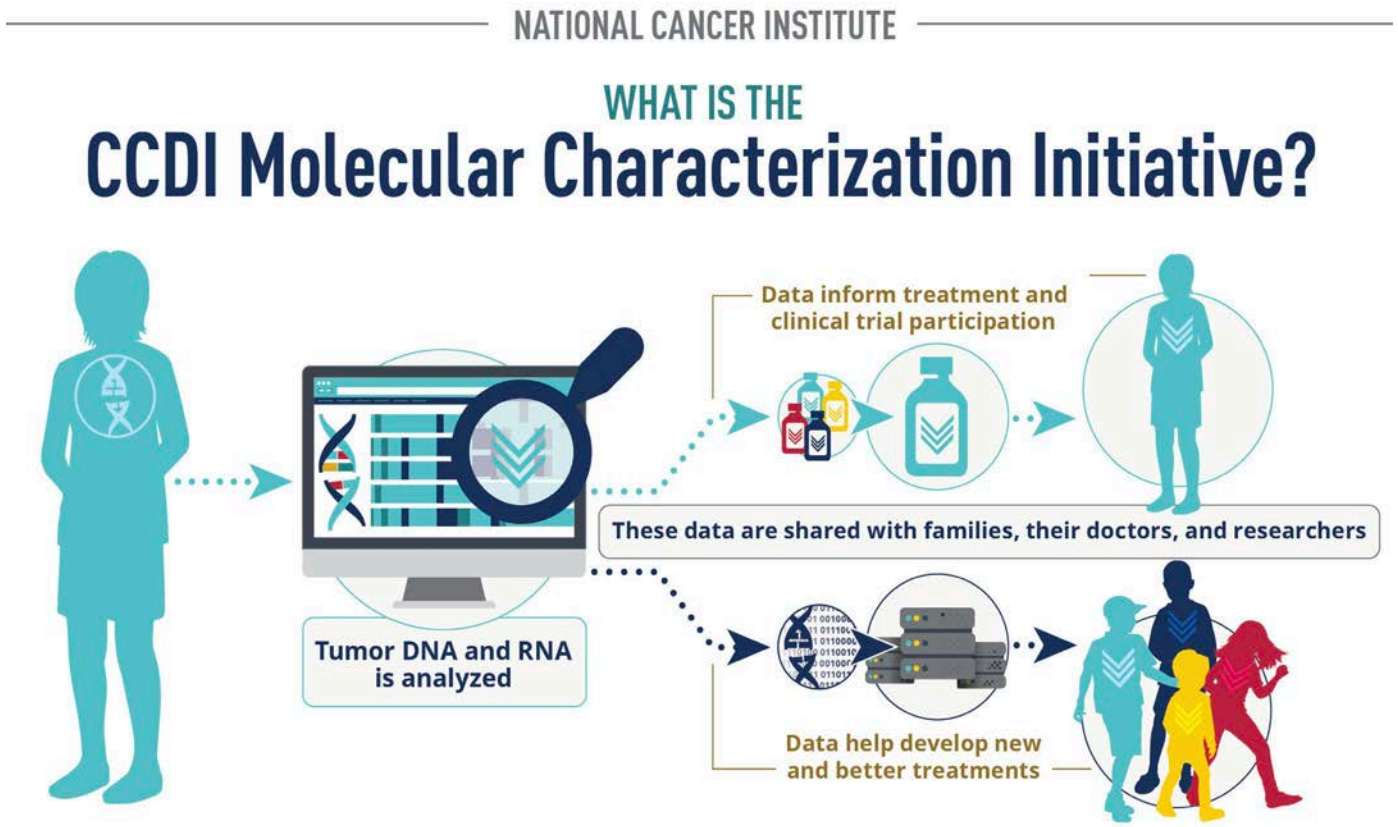
AWS ProSupport: Kevin Sayers

2024 NIH/ODSS Cloud Supplement Program PI Meeting
January 17-18, 2024

Overarching Goals

- Develop cloud-based state-of-the art next generation sequencing analysis (NGS) pipeline based on best practice
- Develop a secure cloud-based comprehensive, genomic integrated browser for visualization and exploration of clinical (patient-centric) and research (cohort summary data) for pediatric cancers
- Support hypothesis generation, publications, and grant applications
- Complimentary to existing browsers (cBioPortal, Genomic Data Commons Data Portal, St. Jude Cloud, and UCSC Treehouse)
- Building on the existing Oncogenomics Clinical NGS pipeline and ClinOmics Portal

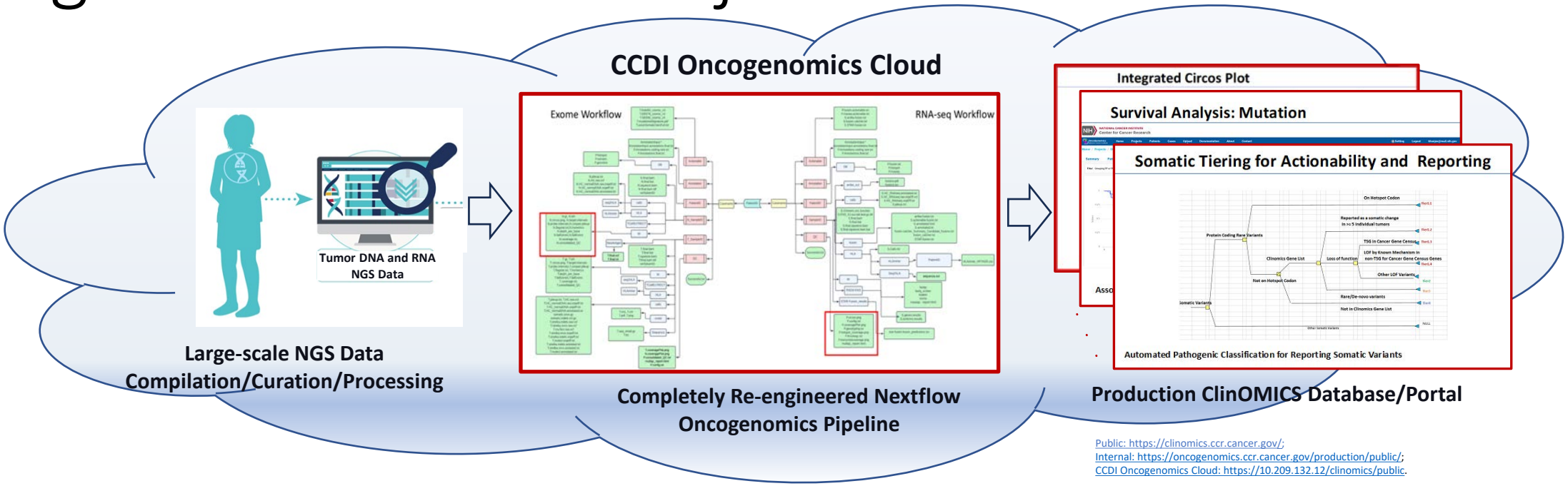
Background: Childhood Cancer Data Initiative (CCDI) Molecular Characterization Initiative (MCI)



Source: https://www.cancer.gov/research/areas/childhood/childhood-cancer-data-initiative

Generate clinical grade next generation sequencing (NGS) and methylation data from every child, adolescent, and young adult diagnosed with childhood cancer, enrolled in Children's Oncology Group (COG) trials nationwide

Childhood Cancer Data Initiative (CCDI) Oncogenomics Cloud Project



Public: https://clinomics.ccr.cancer.gov/;
Internal: https://oncogenomics.ccr.cancer.gov/production/public/;
CCDI Oncogenomics Cloud: https://10.209.132.12/clinomics/public.

Data

Large-scale tumor Whole Transcriptome and tumor/normal Whole Exome **from every child and adolescent young adult**

Oncogenomics NGS Pipeline

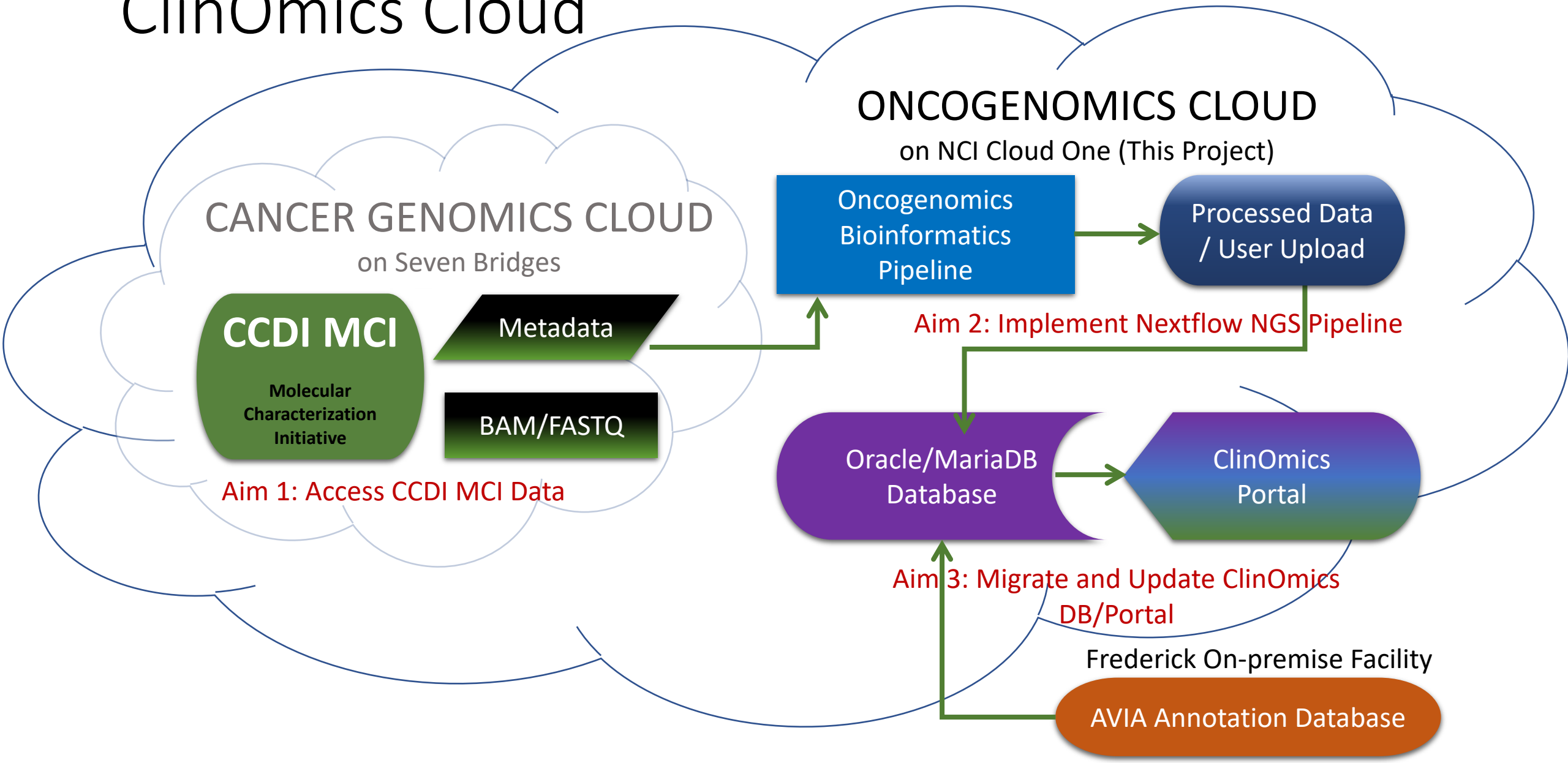
Processing and analyzing large-scale Transcriptomic and WES datasets

ClinOmics DB/Portal

Bring the CCDI MCI and other pediatric cancer genomics data to the fingertips of researchers and clinicians.

ODSS support of NCI data science project for exploratory use of the cloud via the STRIDES Initiative in FY2023, 2024

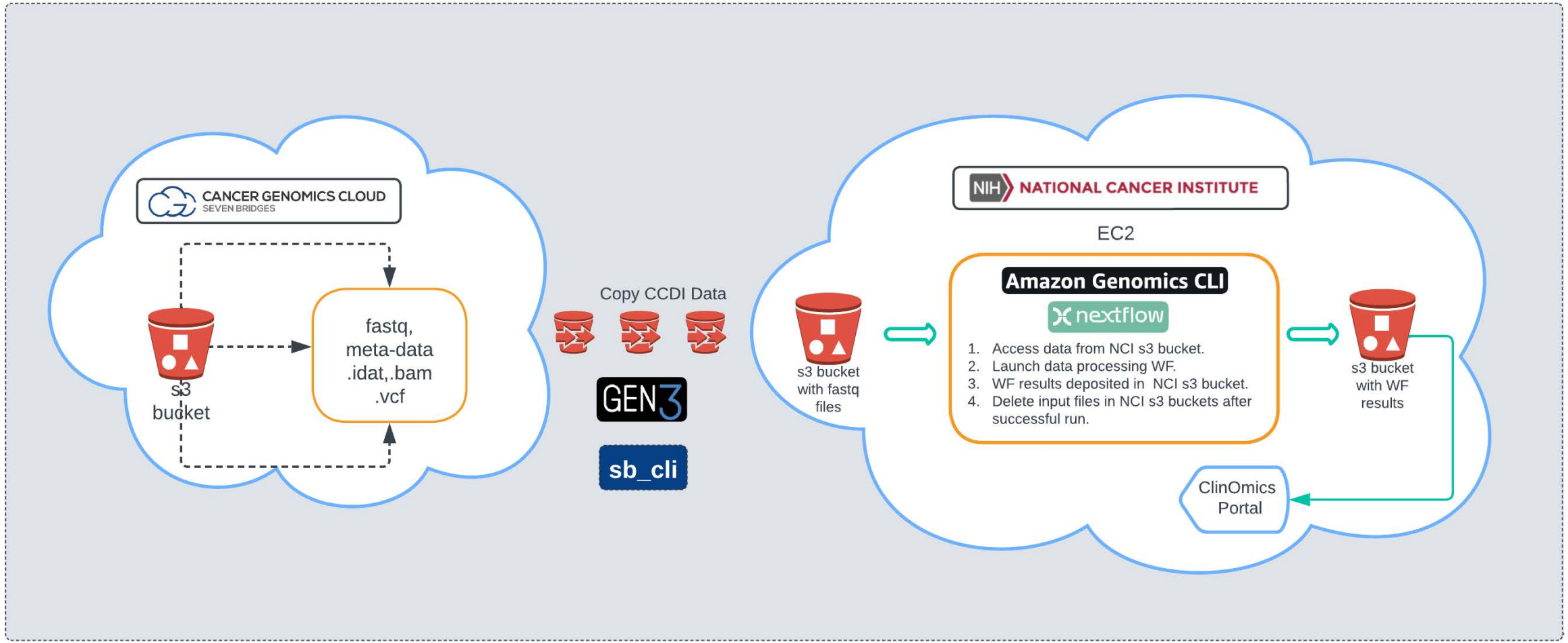
ClinOmics Cloud



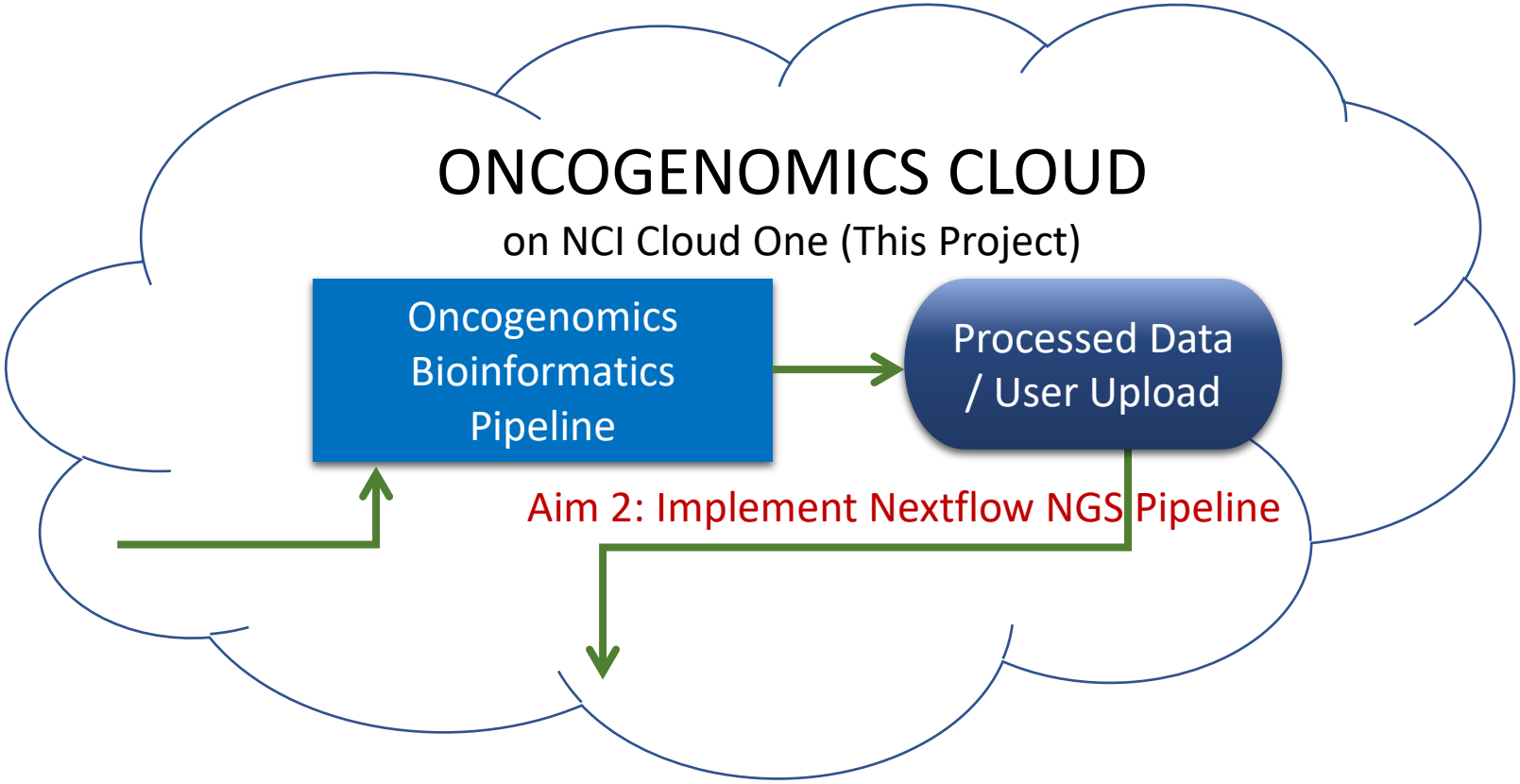
Aim #1: Directly accessing the CCDI Molecular Characterization Initiative (MCI) data in Cancer Genomics Cloud (CGC)

- Explore security and data sharing policies and technological means to directly access the CCDI MCI data in CGC hosted in Seven Bridges.
- Direct access to MCI data in CGC would avoid or reduce data download/egress costs of data analysis via S3 bucket sharing in the AWS cloud.
- The cross-cloud architecture would improve the data transfer, computational, and cost efficiencies of research.
- Pathway to access other childhood cancer genomics data hosted in AWS cloud

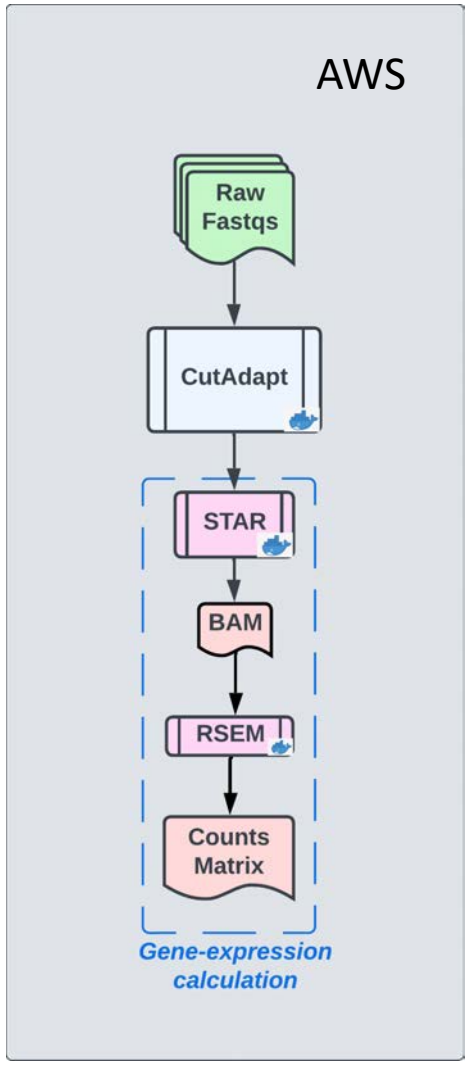
Aim #1 Ongoing: Accessing MCI Data in CGC



Aim #2: Develop and Deploy Oncogenomics NGS Pipeline in the AWS Cloud



Oncogenomics Pipeline – Conversion to Nextflow



Technical Feature/Metric	Snakemake	nextflow
dryrun	✓	✗
"bin" folder	✗	✓
"variable" input	✗	✓
HPC<->Cloud interoperability	✗	✓
s3 fs support	✗	✓
direct GitHub support	✗	✓
inbuilt SLURM support	✗	✓
better regex support	✗	✓
code simplicity	✓	✗

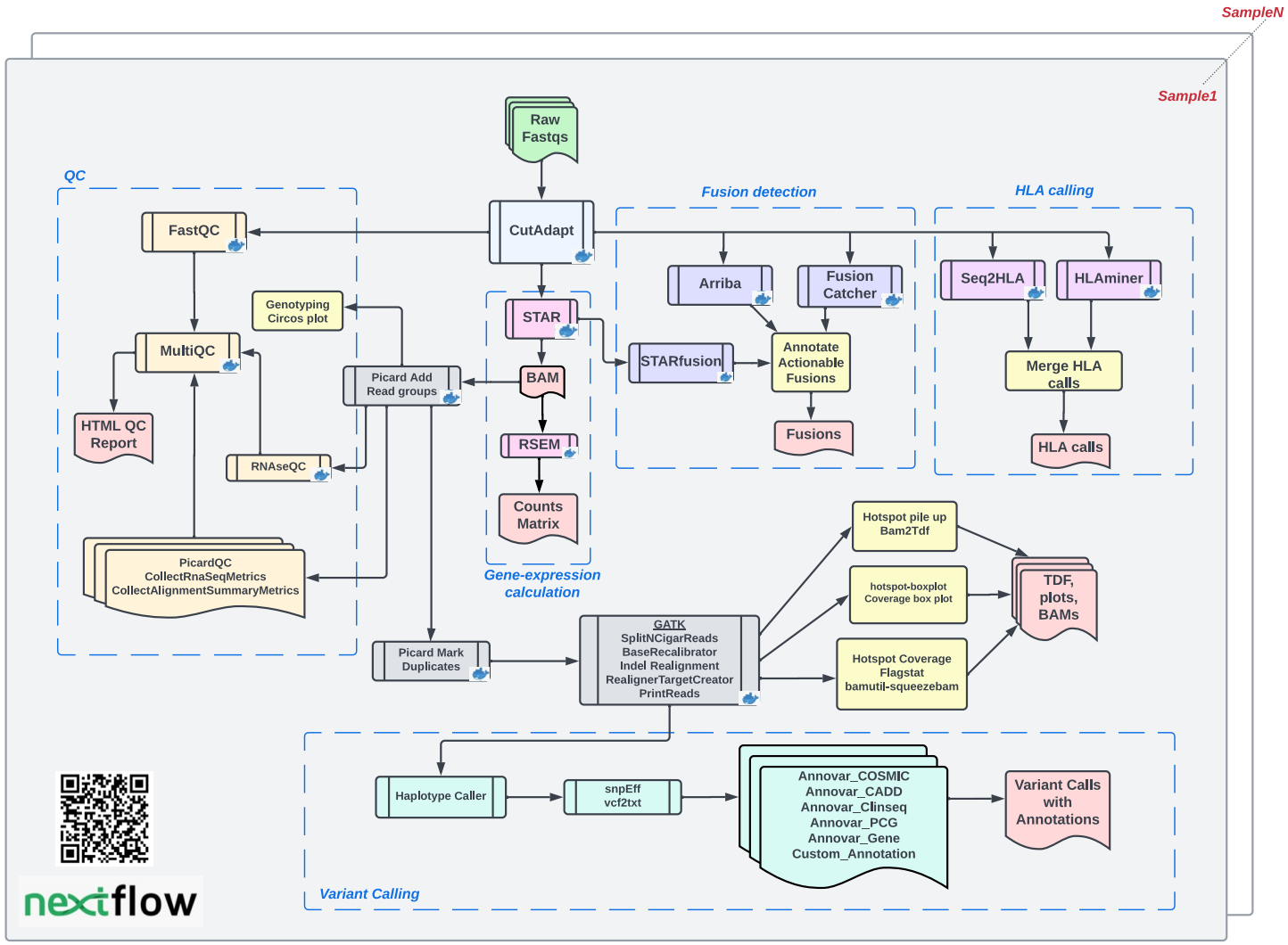
Achieved: Minimal Viable Product (MVP) in Year 1

Re-engineered Oncogenomics RNA-seq Pipeline, deployable on Biowulf, AWS Cloud Computing Platform, and other High-performance environments.

RNASeq

- Gene Expression
- QC
- Fusion Detection
- HLA Calling
- Variant Calling
- HLA

Easy to append new features or switch between genome versions (hg19<->hg38)

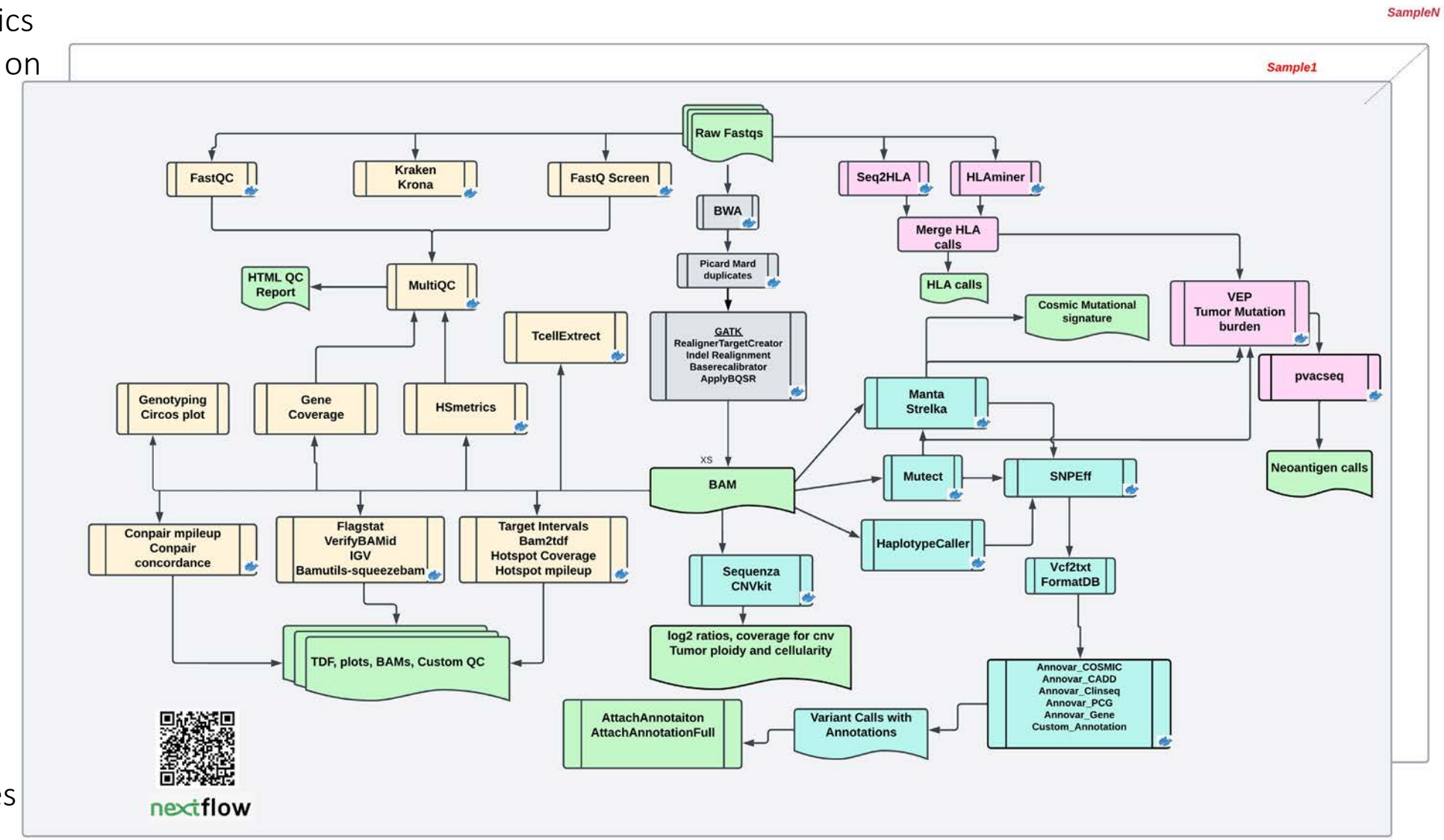


Ongoing: Exome Workflow – Year 1, 2

Re-engineered Oncogenomics Exome Pipeline, deployable on Biowulf, AWS Cloud Computing Platform, and other High-performance environments.

Data generated

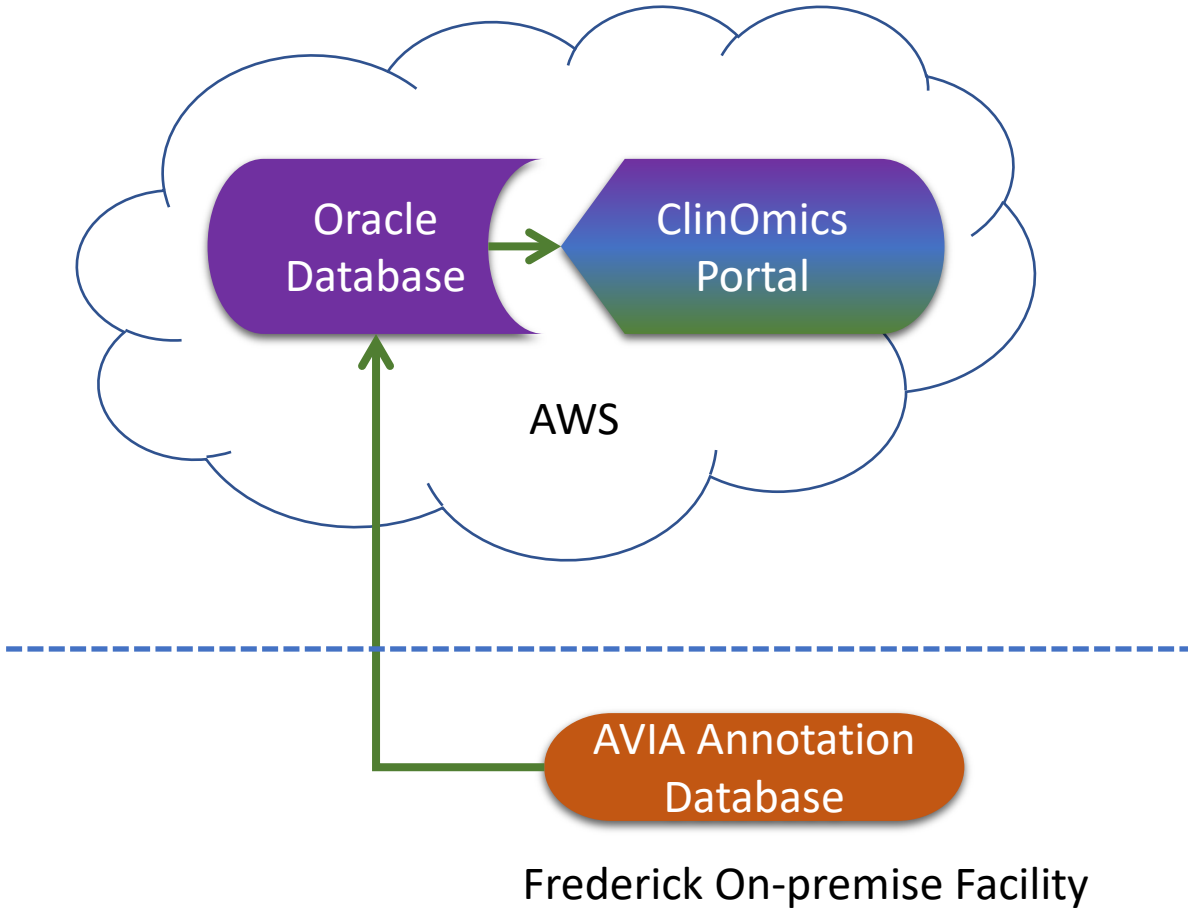
- Actionable Germline
- Actionable Somatic
- HLA
- Copy number variant
- Mutation signature
- Mutational burden
- MSI
- Neoantigen
- Immune infiltrate



Easy to append new features or switch between genome versions (hg19<->hg38)

Aim #3: Migrating DB/Web Portal to AWS

https://clinomics.ccr.cancer.gov/

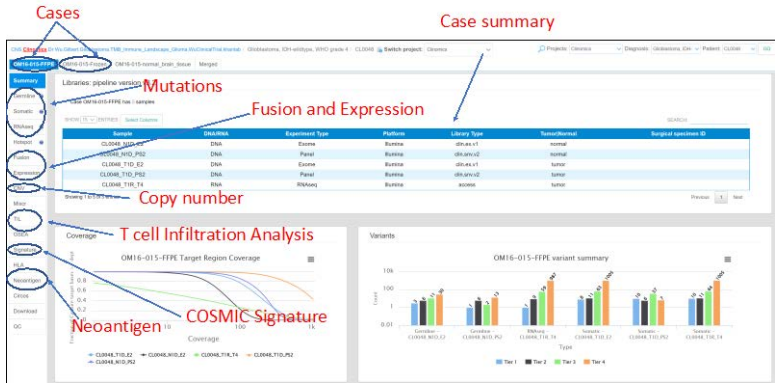


Cloud/On-premise Databases Sync for Up-to-date AVIA Annotations

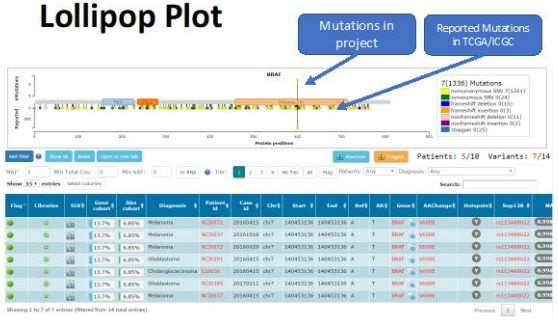
Oncogenomics DB for Clinical and Research Applications

Public: https://clinomics.ccr.cancer.gov/;
Internal: https://oncogenomics.ccr.cancer.gov/production/public/;
Cloud (This Project): https://10.209.132.12/clinomics/public.

Multi-omics individual patient portal

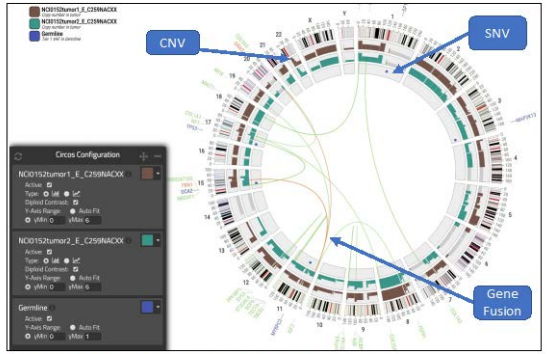


Lollipop Plot



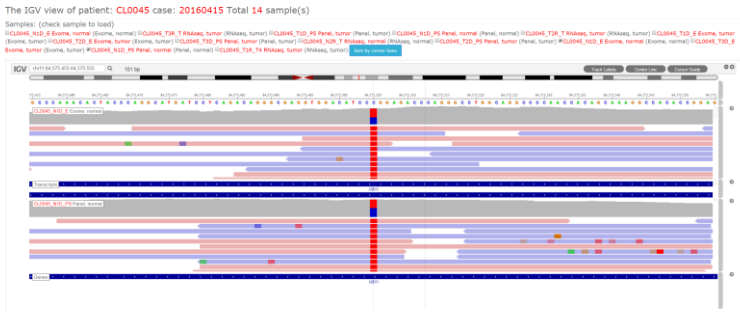
Publications, grant applications, data exploration, biological insights

Integrated Circos Plot



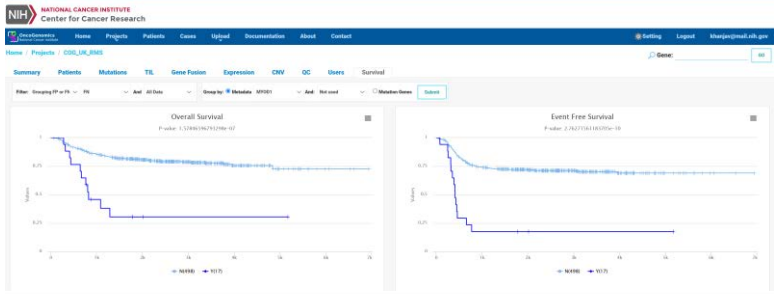
Clinical reports, publications, grant applications

IGV Viewer across all samples for one patient



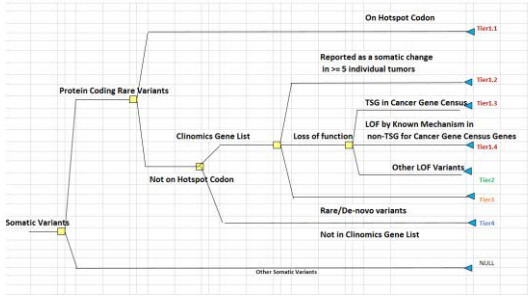
Visual Validation of Variant, Germline vs. Somatic, Tumor evolution

Survival Analysis: Mutation



Association of MYOD1 Mutation with Outcome in Rhabdomyosarcoma

Somatic Tiering for Actionability and Reporting



Automated Pathogenic Classification for Reporting Somatic Variants

Oncogenomics DB for Clinical and Research Applications

Home / Projects / RNAseq_Landscape_Manuscript

Project Level Summary Data

Gene:

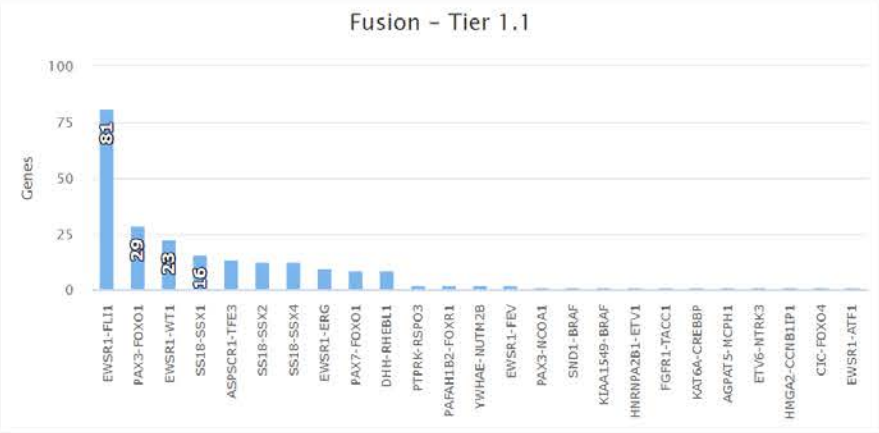
Summary | Patients | Samples | Cases | Mutations | TIL | Fusions | Expression | CNV | Survival | QC | Users

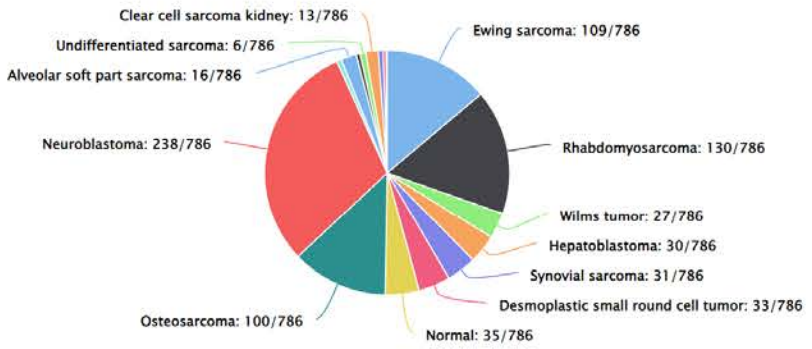
Project ID: 25861 | Version: hg19 | Project Group: 1 | Project Name: RNAseq_Landscape_Manuscript

Patients: 786 | Cases: 294 | Samples: 1696 | Processed Patients: 785 | Processed Cases: 259

Survival: 240 | Exomes: 569 | Panels: 0 | RNAseq: 935 | Whole Genome: 192

Description: RNAseq_Landscape_Manuscript

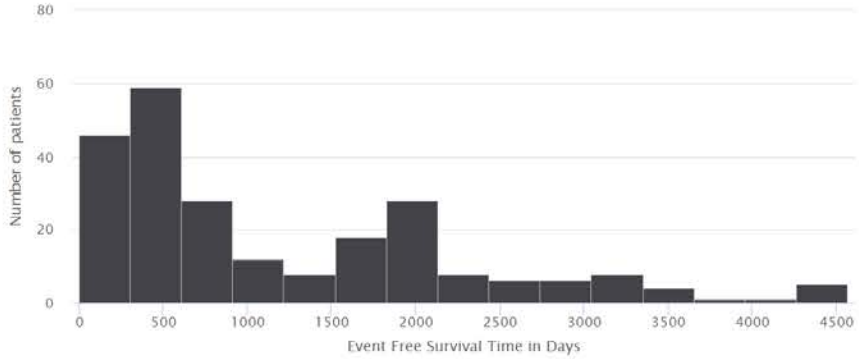


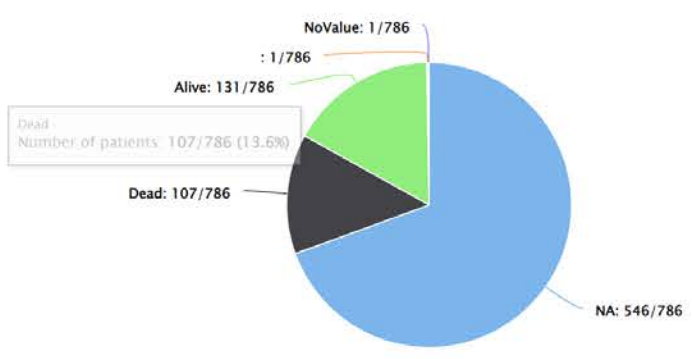


Search:

Diagnosis	Number
Alveolar soft part sarcoma	16
Clear cell sarcoma	3
Clear cell sarcoma kidney	13
Desmoplastic small round cell tumor	33
Ewing like sarcoma	1
Ewing sarcoma	109
Hepatoblastoma	30

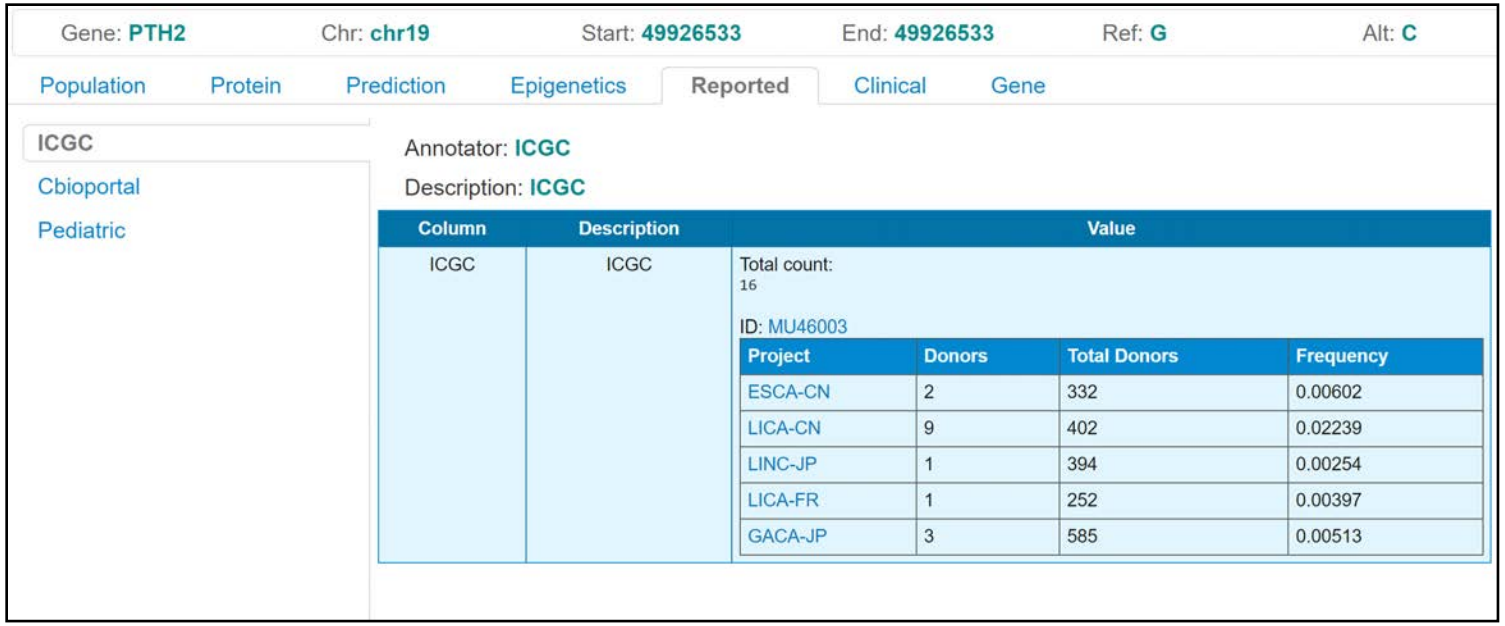
Showing 1 to 17 of 17 entries





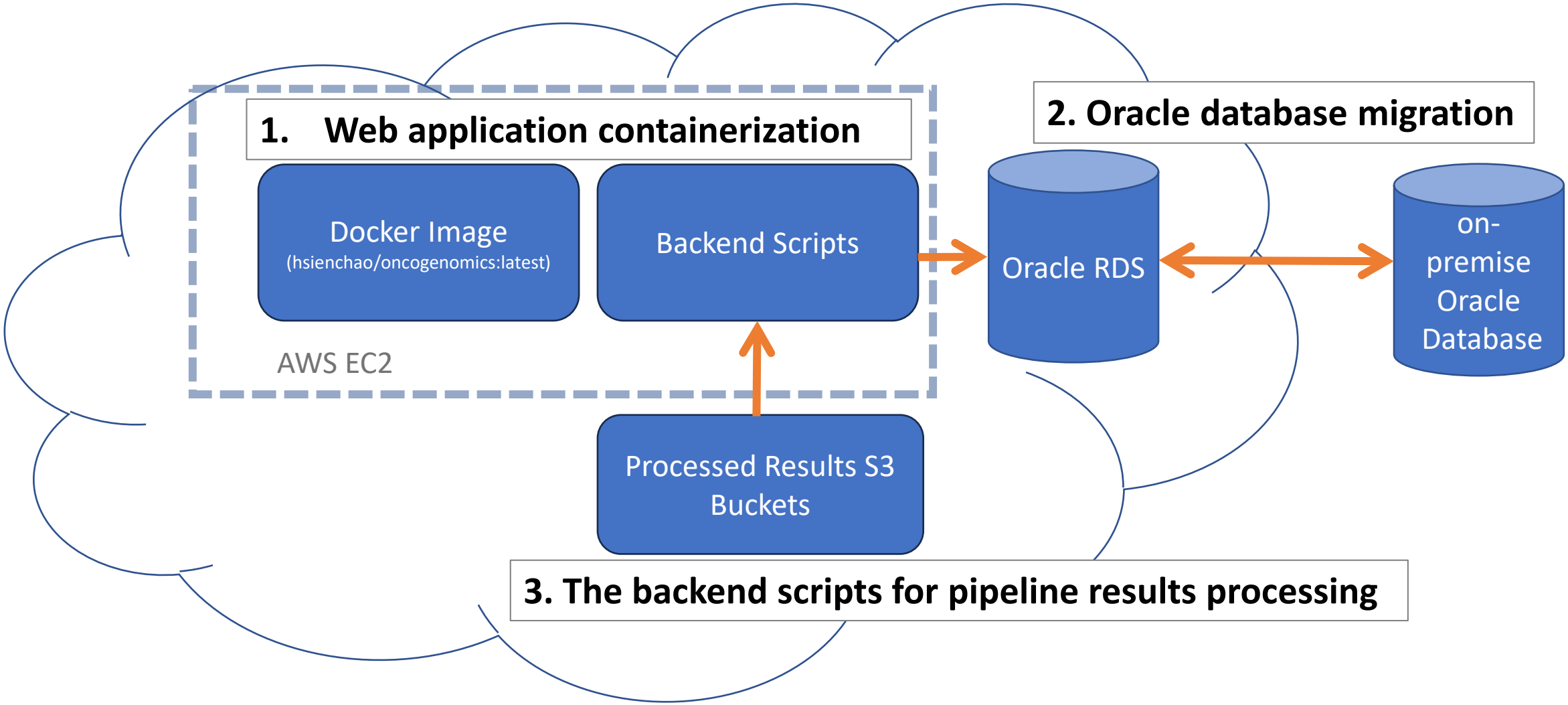
Variant annotation – AVIA (Annotation, Visualization and Impact Analysis)

- Seamless integration with AVIA (https://avia-abcc.ncifcrf.gov/)
- AVIA provides up-to-date customized OpenCRAVAT based annotation
  - Reported database (CBioPortal, ICGC, Genie and PCG)
  - Population frequency
  - Protein annotation
  - Functional prediction
  - Epigenetics information
  - Clinical information
  - Basic gene information



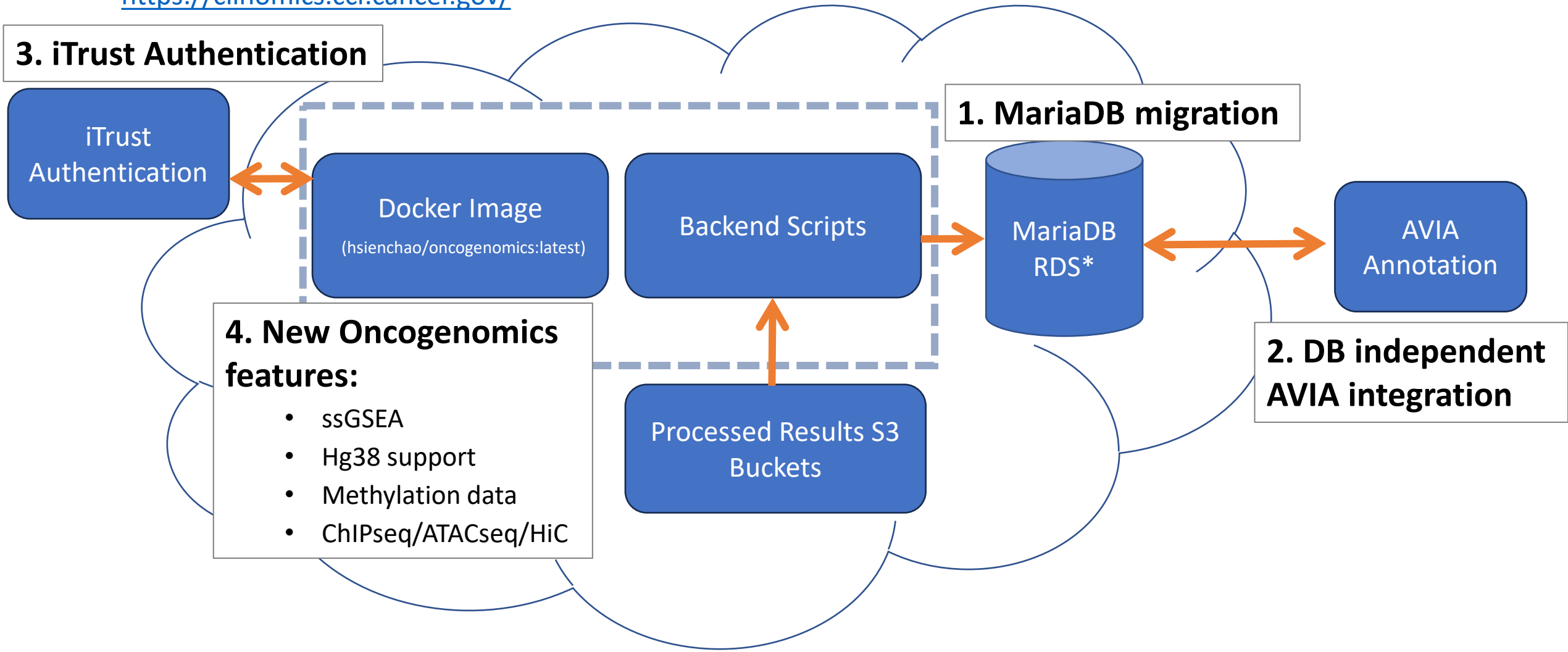
Achieved: Migrating DB/Web Portal to AWS

https://clinomics.ccr.cancer.gov/



Ongoing: Migrating DB/Web Portal to AWS

https://clinomics.ccr.cancer.gov/



Achievements and future work

	Achievements/Products	Planned Future Work
Pipeline	Nextflow RNA- and DNA-seq Pipeline: https://github.com/CCRGeneticsBranch/AWS_POC_MVP_NF	• Add Whole Genome and Methylation workflow modules
	Pipeline Module Docker Containers: https://github.com/CCRGeneticsBranch/Dockers	
Database and Web Portal	The Oncogenomics Database and Web Portal: https://10.209.132.12/clinomics/public	• Incorporate methylation data and tools, including diagnostic classification/ TSNE plots/ deconvolution
	The Oncogenomics Portal Source Code: https://github.com/CCRGeneticsBranch/Oncogenomics_v2	
	The Oncogenomics Portal Docker: https://github.com/CCRGeneticsBranch/Oncogenomics_docker	

Acknowledgements

Genetics Branch

- Patrick Zhao
- Vineela Gangalapudi
- Hsien-chao Chou
- Jun Wei
- Xinyu Wen
- Erica Pehrsson
- … …

ABCS NCI Frederick

- Vishal Koparde
- Anney Che
- Parthav Jailwala
- Uma Mudunuri
- … …

CCDI, CBIIT and AWS

- Subhashini Jagu
- Srujan Boppana
- Krish Seshadri
- Sue Pan
- Kevin Sayers
- … …

ODSS support of NCI data science project for exploratory use of the cloud via the STRIDES Initiative in FY2023, 2024

THANK YOU!